External Data Standards

August 25, 2004
Kathleen Gundry (SAIC)
Hong Dang (AGTI)
Contractors for NCICB

Agenda

- External Standards Review Document
- Types of Standards
- List of Standards in the Review
- Overview of Specific Standards
- Questions and Answers

NCI External Standards Review

- SAIC compiled a review of data standards relevant to the NCI (and caBIG).
- The collection includes both vocabulary/coding standards as well as exchange format standards.
- The report makes recommendations for NCI use of the standards. The table categories them as:
  - A - Recommended for NCI use.
  - B - Recommended for further consideration by NCI.
  - C - Standard is not recommended at this time.

ftp://ftp1.nci.nih.gov/pub/cacore/ExternalStds/

Types of Standards

- Content – controlled vocabularies, ontologies, value lists.
- Information Collection and Format – representation of date and time, PEDro, MIAME.
- Exchange/Transaction – XML, PPI.

Common Demographic/Information Processing and Code Sets

Standard Type	Standard Name	Standard Content	Recommendation
Common Demographic / Information Processing and Code Sets			
Address	FIPS 5-2, Codes for Identification of the States, District of Columbia, and Outlaying Areas of the United States, and Associated Areas	State Names	A
	FIPS 10-4, Countries, Dependencies, Areas of Special Sovereignty, and their Principal Administrative Divisions	Country Names	A
	ISO 3166-1,Country Codes	Country Codes	A
	ISO 11180:1993, Postal Addressing	Address Format	A
	Universal Postal Union	Address Format, State Codes, Country Codes	A
	U.S. Postal Service Postal Addressing Standards	Address Format, State Codes, Street Suffixes, Secondary Unit Designators	B

Common Demographic/Information Processing and Code Sets

Standard Type	Standard Name	Standard Content	Recommendation
Common Demographic / Information Processing and Code Sets			
Language	ISO 639, Codes for representation of language	Language Codes	A
Race and Ethnicity	Office of Management and Budget Directive 15, Standards for the Classification of Federal Data on Race and Ethnicity	Race Identification, Ethnicity Identification	A
Occupation Classification	Bureau of Labor Statistics, Standard Occupational Classification System	Job Activity Classification	A
Vital Statistics	Centers for Disease Control (CDC) National Center for Health Statistics	Birth and death records,Medical records, Interview surveys, Physical exams, Laboratory testing, Marriages and divorces, Fetal death	A
Measurement	HL7 codes for Units, Versions 2.X + (derived from the ISO 2955-83 standard [withdrawn by ISO in 2001] and ANSI X3.50)	Common units of measure, such as Celsius or mg/ml, intended to be combined with a numeric value to accurately express a result	A
	ISO 31, Quantities and units	Individual standards dealing with quantities in space and time, periodic phenomena, mechanics, heat, electricity and magnetism, electromagnetic radiation, chemistry, molecular physics, nuclear physics	A
Information Processing	FIPS 4-2, Representation of Calendar Date for Information Interchange	Means of representing calendar date to facilitate interchange of data among information systems	A
	ISO 8601, Numeric representation of dates and times	Formats for date and time	A

Health-related Vocabulary/Coding Standards

Standard Type	Standard Name	Standard Content	Recommendation
Health-related Vocabulary / Coding Standards			
Health Thesaurus	National Cancer Institute (NCI) Thesaurus	NCI reference terminology and description-logic ontology, providing comprehensive classification and characterization of types of cancer as well as cancer-related diseases, disorders, findings, abnormalities (cellular, molecular, and cytogenetic), gross anatomy, microanatomy, biological processes, genes, gene products, chemicals/drugs, combination therapies, mouse and other experimental models, and other topics	A
Basic Biology	Biological Pathways Exchange (BioPax)	Ontology for pathway information	B
	International Union of Biochemistry and Molecular Biology (IUBMB) and the International Union of Pure and Applied Chemistry (IUPAC)	Controlled vocabulary for nomenclature for biochemistry and molecular biology	A

Health-related Vocabulary/Coding Standards

Standard Type	Standard Name	Standard Content	Recommendation
Health-related Vocabulary / Coding Standards			
Clinical	Common Terminology Criteria for Adverse Events v 3.0 (CTCAE)	Descriptive terminology for Adverse Event reporting	A
	Current Procedural Terminology (CPT) 4	Coding for evaluation and management, anesthesia, surgery, radiology, pathology and laboratory, medicine	B
	Healthcare Common Procedure Coding System (HCPCS)	Healthcare procedures, equipment, and supplies (Level 1) - used for Medicare billing Classification (national level) of physician and non -physician patient care services (Level 2)	B
	International Classification of Diseases for Oncology (ICD-O-3)	Coding for diagnoses of neoplasms - both topography and morphology - includes tumor location, cell type, tumor type, aggressiveness grade	A
	International Classification of Diseases, Clinical Modification (ICD-9-CM)	Classifies diseases, conditions, symptoms, complaints/problems by diagnosis; supplementary classifications include health status, external causes of injury and poisoning, morphology of neoplasms, glossary of mental disorders, drug list numbers, industrial accidents and surgical, diagnostic, and therapeutic procedures	B

Health-related Vocabulary/Coding Standards

Standard Type	Standard Name	Standard Content	Recommendation
Health-related Vocabulary / Coding Standards			
Clinical	International Statistical Classification of Diseases and Related Health Problems (ICD-10)	Collection, processing, classification, and presentation of mortality statistics	A
	Logical Observation Identifiers Names and Codes (LOINC)	Standard test names and codes, descriptive elements for other healthcare areas	A
	Medical Dictionary for Regulatory Activities (MedDRA)	Signs, symptoms, diseases, diagnoses, therapeutic indications, names and qualitative results, surgical and medical procedures, medical/social/family history, adverse event reporting	A
	Systematized Nomenclature of Human and Veterinary Medicine (SNOMED)	Findings/conclusions/assessments, procedures, body structures, function, organisms, substances, physical agents, occupations, social context/demographics, specimens, and other concepts	A

Health-related Vocabulary/Coding Standards

Standard Type	Standard Name	Standard Content	Recommendation
Health-related Vocabulary/Coding Standards			
Genomics	Gene Ontology (GO)	Structured, controlled vocabularies describing gene products used for gene annotations	A
	HUGO Gene Nomenclature Committee (HGNC)	Controlled vocabulary of gene names and symbols for human genes	A
	Mammalian Phenotype Ontology (MP)	Standard vocabulary to describe phenotype data	B
	The Microarray Gene Expression Data (MGED) Society	Comprised of MIAME, MAGE, and the MAGE ontology, a suite of standards for microarray users and developers including an object model, document exchange format, toolkit, and ontology	A
	Mouse Anatomy (adult – MA, and development – EMAP)	Ontologies used to annotate gene products	A
	Taxonomy	National Center for Biotechnology Information (NCBI) taxonomy of organism names represented in genetic databases	A
Drug Identification	National Drug File Reference Terminology (NDF-RT)	Drug classes, active ingredients (chemical structure), mechanics of action, physiologic effect, pharmacokinetics, therapeutic intent, commercial/clinical drug identification	A
	RxNorm Clinical Drug Vocabulary	Ingredients, drug components, drug formulations, drug strength representation, drug name synonyms, dosage forms	A

Health-related Transaction Standards and Models

Standard Type	Standard Name	Standard Content	Recommendation
Health-related Transaction Standards and Models			
Imaging	Digital Imaging and Communications in Medicine	Standard method for the transmission of medical images and their associated information	A
Basic Biology	Systems Biology Markup Language (SBML)	XML exchange format for exchange of biochemical network models	A
	CellML	XML-based language for describing and exchanging models of cellular and subcellular processes	A
Clinical	Health Level Seven (HL7)	Patient tracking, scheduling, orders, results, clinical observations, billing, medical records, patient referral, patient care	A
	Clinical Data Interchange Standards Consortium (CDISC)	Clinical trials data - general data (study name, protocol name, measurement units), study metadata (code lists), administrative data, reference data (lab normal ranges), clinical data	A
	North American Association of Central Cancer Registries, Inc. (NAACCRz, Inc.)	Demographic, tumor and staging, treatment and follow-up	A
Genomics / Proteomics	Macromolecular Structure (Mms)	Specification for a data model and interface for exchange of macromolecular structure information	B
	PEDRo	Data model implemented in SQL and XML to support proteomics research	B
	Protein-Protein Interaction (PPI)	Data exchange format designed to bridge different formats of protein interaction databases	B
	Tissue Microarray (TMA)	Data exchange specification for tissue microarray data	A

Standards Review Bodies

Standard Type	Standard Name	Standard Content
Standards Review Bodies		
	Centers for Disease Control Public Health Information Network (PHIN)	Vocabulary and messaging standards; standards for data display and entry; standards for data transmission and management; implementation of applications and databases to support the adopted data standards
	Consolidated Health Informatics Initiative (CHI)	Portfolio of existing clinical vocabularies and messaging standards enabling federal agencies to build interoperable federal health data systems
	Food and Drug Administration, Center for Drug Evaluation and Research (CDER)	Compilation of standardized nomenclature monographs that have been reviewed and approved by the CDER Nomenclature Standards Committee (NSC)
	National Council on Vital Health Statistics (NCVHS)	Advises the government on recommended standards for adoption in the health care sector

Overview of Reviewed External Biology Standards

System	TAX MP	TMA	*Tissue Array*
Organic	MA/EMAP MPATH		*Metabolomic*
Cellular	SBML/CellML/BioPAX GO	PEDRo PSI-MS PPI *PSI-OM/-ML /-Ont/MIAPE*	*Proteomic*
Molecular	PPI/Mms HGNC BSA/GM/SO IUPAC-IUBMB	MAGE MGED Ont MIAME Tox-MIAME	*Gene Exp.*

Overview: Issues Associated with Standards

- Fulfilling a need/niche, timing.
- Community participation – create/use.
- Ease of use, availability of software tools.
- Accommodate change.
- New standards are being created.

Gene Ontology (www.geneontology.org)

- 3 Structured vocabularies: biological Process, molecular function, cellular component.
- Used by most, if not all, genome databases.
- Sample apps: GO browsers (Amigo, QuickGO…), EASE, GoMiner, FatiGO, GOSurfer…
- Main issue: regular updates (monthly).

Basic Biology Nomenclatures

(http://www.chem.qmw.ac.uk/iupac/jcbn/)

- IUPAC-IUBMB: Biochemical nomenclature recommendations – amino acids, nucleic acids, lipids, carbohydrates, enzymes, tocopherols …
- HGNC: 19775 approved gene symbols, 21574 aliases, and 4363 withdrawn symbols, as of Friday August 13 20:58:50 2004.
- NCBI Taxonomy: 118,051 species, 176,972 taxa; used by all genome databases, not a phylogenetic or taxonomic authority.

Microarray Gene Expression Data (MGED) Society

(http://www.mged.org/)

MAGE (MicroArray Gene Expression):

- MAGE -OM: the object model
- MAGE-ML: the exchange format
- MAGE-stk: the software toolkits

MIAME (Minimum Information About a Microarray Experiment):

- Compliant data repositories and databases: ArrayExpress, GEO, CEBS, SMD…
- Extensions: tox-MIAME…

MGED ontology: controlled vocabulary for experiment annotation.

Proteomics

Proteomics Experiment Data Repository (PEDRo) (http://pedro.man.ac.uk/)

- Data models: mass spectrometry, liquid chromatography, gel electrophoresis…
- Extensible: COGENE, MAGE-ML, CHIME, Ontology…
- Software.

HUPO Proteomics Standards Initiative (PSI) (http://psidev.sourceforge.net/)

- Protein-Protein Interaction (PPI): data model and exchange format – PSI MI (molecular interaction) XML; adoption by major interaction databases such as BIND, DIP, MINT…
- PSI-MS: data model and exchange format for mass spec.
- PSI-OM, PSI-ML, MIAPE, PSI-Ont are being developed modeling the development of MGED standards.

System Biology

Systems Biology Markup Language (SBML)
(http://sbml.org/index.psp)

- Data representation and exchange format for quantitative models of biochemical reaction, regulatory networks.
- Adopted by > 60 modeling software systems as of 08/2004.

CellML
(http://www.cellml.org/public/news/index.html)

- Format for storing and exchange of mathematical models of cellular systems.
- Related projects AnatML and FieldML.
- Working with SBML to ensure compatibility.

Biological Pathways Exchange (BioPAX)
(http://www.biopax.org/index.html)

- Developing pathway data exchange format in the form of an ontology.

The Tissue Microarray (TMA) Data Exchange (http://www.biomedcentral.com/1472-6947/3/5)

- NCI and API sponsored standard.
- XML specification of data elements relevant to tissue microarray.
- Available as CDEs.

Genome and Sequence Standards

Sequence Ontology (SO)
(http://song.sourceforge.net/))

- Part of the Open Biological Ontologies (OBO).
- Provide structured controlled vocabulary for sequence features.

Biomolecular Sequence Analysis (BSA) Specification
(http://www.omg.org/technology/documents/formal/biomolecular_sequence.htm)

- Data model specification for representing bio-sequence objects and analysis.
- OMG spec, but not used by major databases to our knowledge.

Genome Maps (GM) specification
(http://www.omg.org/technology/documents/formal/genomic_maps.htm)

- Data model specification for representing genomic maps.
- OMG spec, but not used by major databases to our knowledge.

Organ and System Level Standards

Mouse Anatomy (MA)
(http://www.informatics.jax.org/searches/anatdict_form.shtml)

A

- 28 controlled vocabularies corresponding to the stages of embryonic and postnatal development of the mouse.
- Potential overlap with other anatomy vocabularies.

Mouse Pathology (MPATH)
(http://eulep.anat.cam.ac.uk/))

C

- Structured vocabulary to annotate histopathology images in Pathbase.
- Potential overlap with other vocabularies, such as clinical.

Mammalian Phenotype (MP) ontology
(http://www.informatics.jax.org/searches/MP_form.shtml)

C

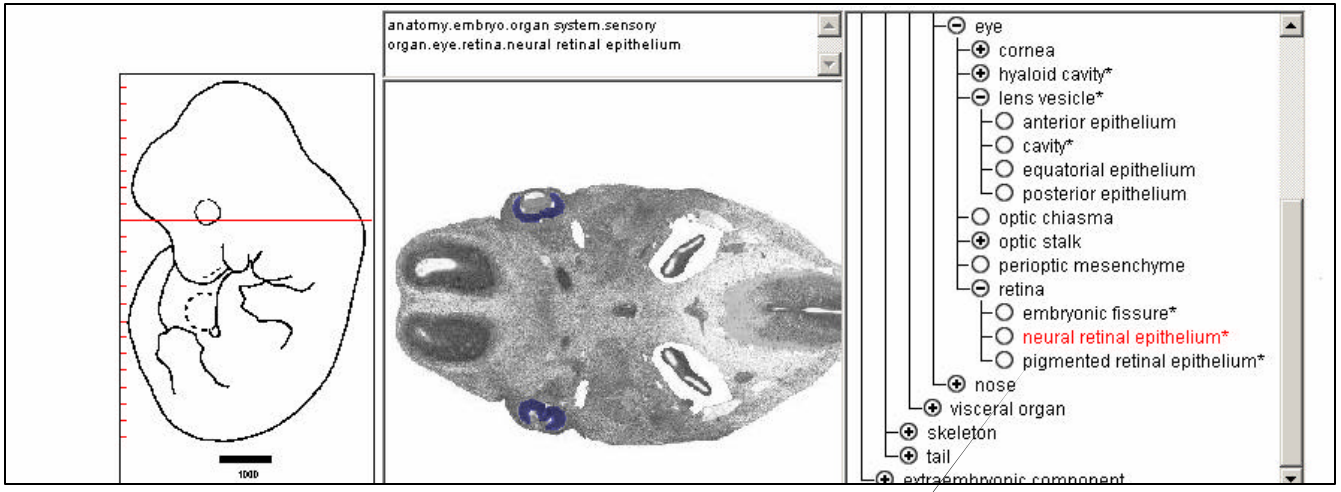
- Structured vocabulary for annotating phenotype data.
- Under development; potential overlap with other vocabularies.

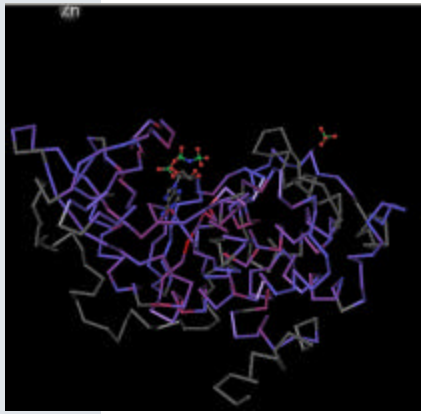
Macromolecular Structure (Mms)

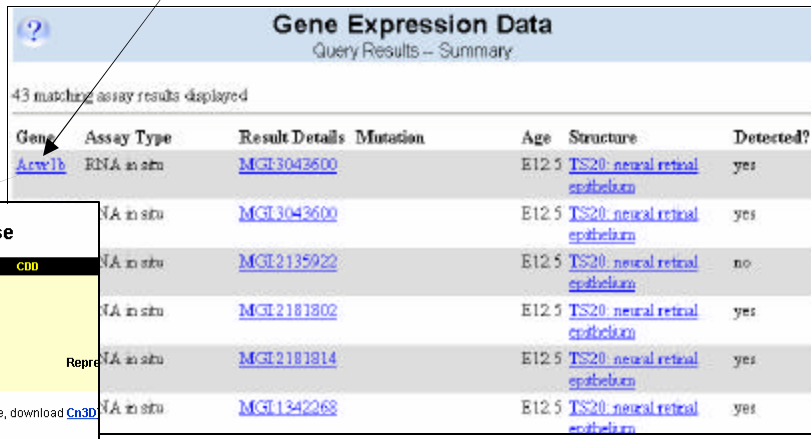
(http://openmms.sdsc.edu/)

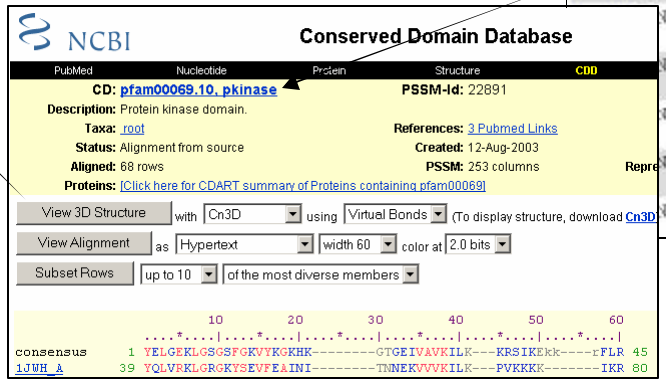
- Detailed data model definition for information related to macromolecular structures.
- Authoritative: NIST, OMG, RCSB (PDB).
- Available software toolkit (OpenMMS) for parsing structural information from mmCIF files.

Example of Data Integration: MA & GXD









Questions and Answers